Új integrált magyar morfológiai elemző

Novák Attila[1,2]

[1] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport,
[2] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar,
1083 Budapest, Práter utca 50/a
e-mail:{novak.attila}@itk.ppke.hu

Kivonat A morfológiai elemző a magyar nyelvtechnológiai alkalmazásokban kulcsszerepet játszik, hiszen minden nyelvi feldolgozási lánc első lépései között szerepel, így minden ezen a szinten keletkező hiba tovább terjed. Kulcsfontosságú tehát egy olyan morfológiai elemző és keretrendszer létrehozása, amely szabadon hozzáférhető, könnyen bővíthető és módosítható egy-egy adott felhasználási terület sajátos igényei szerint, és nem utolsó sorban jó minőségű elemzésre képes. Cikkünk célja egy ilyen morfológiai elemzőrendszer tervezési szempontjainak és a már létező morfológiai elemzőkre alapuló implementációjának bemutatása.

1. Bevezetés

Az MTA INFRA2 pályázatának keretein belül megvalósuló nyílt, integrált magyar nyelvtechnológiai kutatási infrastruktúra fejlesztésének célja egy olyan nyílt forrású, szabadon hozzáférhető nyelvtechnológiai infrastruktúra fejlesztése, melynek elemei a magyar nyelv gépi elemzésének alapvető eszközeit tartalmazzák egy integrált, szabványos keretben. A morfológiai elemző az infrastruktúra fejlesztésének központi eleme. A cél egy szabadon elérhető, az eddigi eszközök tudását szintetizáló, gyors és testre szabható elemző eszköz és a hozzá kapcsolódó komplex fejlesztőkörnyezet létrehozása.

2. Morfológiai elemzők magyar nyelvre

A jelenleg magyar nyelvre elérhető morfológiai elemzők (Humor [5,7], Xerox és Hunmorph/morphdb.hu [9]), illetve a hunmorph-foma[3] elemző minősége eltér, más-más nyelvi jelenséget tudnak jól kezelni, illetve tőtáruk is különböző, ezért különböző szókincset fednek le. Ugyanakkor az egyes elemzők leírásánál alkalmazott formalizmus is jelentősen eltér. Emellett az említett erőforrások jelentősen különböznek a leírás olvashatósága, illetve karbantarthatósága, a lefedett szókincs bővíthetősége és a forrás hozzáférhetősége szempontjából, illetve abból a szempontból is, hogy az erőforrás fejlesztői mennyire érhetőek el. Az utóbbi két szempont miatt a Xerox magyar elemzőjének forrásként való felhasználása

[3] http://freecode.com/projects/hunmorph-foma

nem jöhet szóba. A hunmorph-foma elemző a leírás olvashatósága, illetve karbantarthatósága, módosíthatósága, valamint a lefedett szókincs bővíthetősége szempontjából messze elmarad a Humor és a morphdb.hu alapú erőforrások mögött. Ez a leírás ugyanis nem nyelvtanon alapul, bővíteni kizárólag analógiás alapon lehet, a felvenni kívánt új szóval azonos morfológiai viselkedésű szó leírásának lemásolásával. Nem is beszélve a leírásban, illetve a paradigmákban levő esetleges hibák javításáról. Ezen kívül a gitorius.org lekapcsolásával a forrása pillanatnyilag nem hozzáférhető, és a fejlesztő elérhetősége is kérdéses. Mindezek miatt ennek az erőforrásnak a forrásként való használatáról is lemondtunk.

A morphdb.hu-n alapuló hunmorph (ocamorph) és jmorph elemzők nagy előnye, hogy nyílt forráskódúak, és maga a morfológiai adatbázis egyrészt nyelvtanalapú, így jól bővíthető, javítható, a forrása alapján értelmezhető, másrészt teljesen szabadon felhasználható és módosítható. A morphdb.hu és a morfológiai leírás alapjául szolgáló hunlex eszköz hátránya ugyanakkor, hogy ezek fejlesztése sok évvel ezelőtt leállt, a dokumentáció hiányos, és bár az egykori vezető fejlesztő, Trón Viktor kifejezte együttműködési készségét a projekt résztvevőivel, egyrészt csak nagyon korlátozott időben tud a rendelkezésünkre állni, másrészt a rendszer nem dokumentált tulajdonságai és a befejezetlen fejlesztések részleteivel kapcsolatban nem nagyon tudott segíteni, mert az évek folyamán a kérdéses tudás feledésbe merült. Ugyancsak nem könnyíti meg a helyzetet, hogy a hunlex implementációja OCaml nyelven íródott, és nem áll rendelkezésünkre ezen a nyelven kompetens programozó. Ugyancsak OCaml nyelven íródott az ocamorph elemző, ezzel szemben a jmorph Java alapú. A két elemzőeszköz azonban különbözik a bennük implementált elemzőalgoritmus, különösképpen az összetételi konstrukciók kezelése szempontjából. Ugyanakkor a különbségek nem dokumentáltak, pusztán az eszközök viselkedésének vagy forráskódjának tanulmányozása révén tárhatók fel.

A Humor elemző alapjául szolgáló morfológiai adatbázis forrása az ebben a cikkben említett projektum elindulásáig nem volt szabadon hozzáférhető, és maga a Humor elemző is zárt forráskódú. Ugyanakkor ez az erőforrás is nyelvtanalapú, így jól bővíthető, javítható, illetve viszonylag jól dokumentált. Az említett erőforrások közül egyedüliként a fejlesztő elérhető, és a rendszer fejlesztése annak létrehozása óta folyamatos. A Humor morfológiai adatbázis jellemzői az 1. táblázatban láthatók.

3. Az elemzők fedésének kiértékelése

Első lépésként a további fejlesztés szempontjából szóba jövő elemzőknek (Humor, ocamorph, jmorph) a részletes minőségi kiértékelése, és kritikai összehasonlítása volt a feladat egy nagy szöveges korpuszból készült gyakorisági lista segítségével. Egy 4 milliárd szavas, nagyrészt webről letöltött szövegből készült, elemi tokenizálóval tokenizált, nem szűrt 35 millió szavas gyakorisági listát elemeztettünk a Humor, az ocamorph és a jmorph elemzőkkel. Az ocamorph elemző produktív összetettszó-elemző üzemmódját bekapcsolva sajnos az input szavak egy részére végtelen ciklusba került, ezért egy idő után kikapcsoltuk ezt az üzemmódot.

1. táblázat. A magyar Humor morfológia jellemzői

tőlexikon		**lemmák/lexémák**	**allomorfok**
általános szókincs		**95811**	**141718**
	eredeti lexikon kibővítve	75132	105473
	zárt tőosztályok (névmások, számnevek, stb.)	744	3675
	szótárakból és korpuszokból	19935	32570
terminológiai lexikonok		**110129**	**178324**
	Földrajzi és személynevek	40262	
	Nukleáris terminológia	911	
	Gazdaság/adminisztráció	4736	
	Angol	1920	
	Orvosi	40813	
	Katonai	21487	
összes		**205940**	**320042**
	ebből összetett	89415	126728
	sokmorfémás/toldalékolt		7720

toldaléklexikon		**lexémák**	**allomorfok**
	összes	**283**	**12041**
	sokmorfémás		10959

szabály file-ok	**műveletek**	**szabályok**	**sorok**
tőszabályfile			
	45 deklaráció	520 szabály	2074 sor
	596 allomorfgeneráló művelet	220 tő allomorfiaszabály	
toldalékszabályfile		50 szabály	233 sor
	86 allomorfgeneráló művelet	34 allomorfiaszabály	

	állapotok	**átmenetek**	**flagek**
szónyelvtan-automata			
	47 állapot	602 átmenet	20 flag

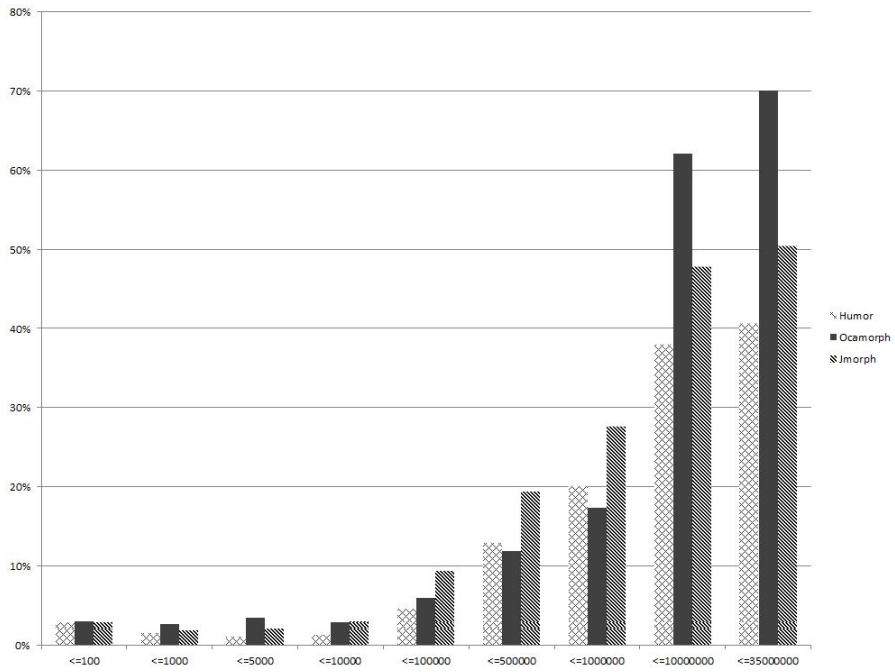
	kategóriák	**tulajdonságok**
a jegyek és szónyelvtan-kategóriák kódolásának definíciója		
	102 szónyelvtan kategória	102 vektorkódolt jegy
		187 mátrixkódolt jegy

Emellett mivel a morphdb.hu nyelvtanban az összetett szavak szerkezetét leíró konstrukció egyszerűen tetszőleges nominális (főnévi, melléknévi, számnévi) tövek tetszőleges sorrendben tetszőleges számban való megjelenését megengedi

(tehát pl. a *tevepirosnegyven* szó helyes számára), ezért sokkal több értelmetlen elemzést ad a produktív elemzés bekapcsolása esetén, mint a másik két elemző. A futtatás során a 2. táblázatban látható eredményeket kaptuk az egyes elemzőkre. A 1. ábrán látható, hogy az 500000. és a 10000000. szó közötti régióban az ocamorph adta a legnagyobb lefedést, a többi régióban a Humor.

2. táblázat. A nem elemzett szavak száma a 35 millió szóalak közül az egyes elemzők esetén.

Elemzőrendszer	Nem elemzett szavak száma
Humor	13 754 680
ocamorph	23 248 165
jmorph	17 152 815



1. ábra. A három elemző által nem elemzett szavak aránya gyakorisági rangsor egyes régióiban

A jmorph és az ocamorph ugyanazon lexikai adatbázis feltehetőleg különböző változatainak felhasználásával készült. Abban, hogy jelentősen különböző eredményt adnak, az elemzőkben implementált algoritmusok különbsége is szerepet játszik. A gyakori régióban előforduló hiányok elsősorban tokenizálási, központozási és helyesírási hibákból adódnak.

4. Az új morfológiai elemzőrendszer

A kifejlesztendő morfológiai elemzőrendszer felépítése három rétegből áll. Az első a felhasználó, nyelvész szakértő által olvasható morfológiai forrásadatbázis, azaz a tőtár és a morfo(fono)lógiai nyelvtan. A második réteg egy forrásadatbázis-konverter, mely a harmadik réteg számára szükséges erőforrásokat állítja elő az első rétegből. A harmadik réteg pedig maga az elemzőfuttatási keretrendszer.

Mind az ocamporh és jmorph elemzők , mind a Humor elemző által használt adatbázisok egy ugyanígy hármas tagolású rendszerben jönnek létre. A projekt jelenlegi fázisában a Humor nyelvtan és a [6] cikkben leírt véges állapotú nyelvleírást előállító konverter segítségével generáljuk az elemző adatbázisát.

4.1. A tőtár

A forrásadatbázis a Humor és morphdb.hu adatbázisok tőadatbázisának és morfológiai szabályrendszerének szintéziseként áll elő. Támogatja a pragmatikai, nyelvhasználati, szemantikai és morfológiai jellemzők rögzíthetőségét és kezelését. Pragmatikai jellemzők alatt a különböző stílusminősítő jegyeket, a helyesírási normától való eltérést jelző tulajdonságokat, illetve gyakorisági információkat értünk. Ezek a korábbi elemzők egyikében sem voltak egyszerre jelen. A forrásadatbázis a szemantikai jegyek, az ontológiai besorolás mellett a tematikus/vonzatkereteket és az időaspektusra vonatkozó jegyeket is tartalmazza. A morfológiai jegyek a Humor és a morphdb.hu lexikonok kategória-rendszerének egyesítésével lettek meghatározva, a kettő uniójaként, az esetleges ütközések és ellentmondások feloldása mellett. Bár a Humor adatbázis eleve tartalmazott az elemzésben meg nem jelenített szemantikai jellegű címkéket, ezeket a projektum során disztribúciós szemantikai modellek felhasználásával kibővítjük és ellenőrizzük [8].

A Humor adatbázisában szereplő mintegy 200000 lemmán túl a morphdb.hu adatbázisa kb. 13000 új lemmát tartalmaz, bár ezek egy része a helyesírási normának nem megfelelő alak. Jelenleg folyamatban van ennek a 13000 alaknak az ellenőrzése és a helyesírási normáknak nem megfelelő alakok leképezése a megfelelő helyes alakokra. Ehhez a Humor tőadatbázisából és az Osiris Helyesírás [3] szótári részében szereplő szavakból és többszavas kifejezésekből épített listán az A* algoritmust [2] futtatva és hibamodellként tévesztési mátrixot definiálva rangsorolt javítási javaslatokat generáltunk, amelyeket a kézi ellenőrzés támogatására használunk.

4.2. A kategória-rendszer

A főkategóriák mellett alkategóriákat is bevezetünk, melyek egyrészt a Humor/morphdb.hu rendszerben szereplő, de a végső kategória-rendszer főkategóriái közé nem kerülő kategóriák, másrészt az egyéb szemantikailag vagy morfoszintaktikailag releváns kategóriák. Továbbá a szokásos morfológiai jegyek: inflexiók, képzők, szóösszetételi határok és típusok is a két elemző adatbázisának egyesítésével kerülnek bevezetésre. A létrejött elemző fonológiai jellemzők kezelésére is alkalmas. Ez magában foglalja a szóalakok CV-vázának meghatározását (a rövid-hosszú szegmentumok opcionális megkülönböztetésével). Kezeljük továbbá a a felszíni alakból automatikusan nem levezethető kiejtéseket, illetve kiejtésváltozatokat és a nyelvjárási és szociolektális megkülönböztetéseket.

Ehhez áttekintettük a morphdb.hu nyelvtan hunlex nyelven megírt forrását és megkerestük azokat a pontokat, ahol a Humor leírásba átemelhető eltérések vannak. A morphdb.hu adatbázisban a határozószavak számos alosztályra vannak osztva. Ezt az osztályozást disztribúciós módszerrel validáltuk és kiterjesztettük. Ehhez a nyelvtechnológiai kutatások egyik kurrens módszerét, a neurális hálózattal létrehozott folytonos vektoros reprezentációkból álló modellt alkalmaztuk (*word embedding*). Ez a módszer nyers szöveges korpuszból szemantikai és grammatikai információk kinyerésére alkalmazható úgy, hogy az eredményül kapott modellben a lexikai elemek egy valós vektortér egyes pontjai, melyek konzisztensen helyezkednek el az adott térben, azaz a hasonló disztribúciójú, egymáshoz szemantikailag, szintaktikailag és/vagy morfológiailag hasonló szavak egymáshoz közel, a jelentésben eltérő elemek egymástól távol esnek. A modell létrehozásához a word2vec[4] eszközt használtuk a [8] cikkben ismertetett módon. Az eredmények számos hibára rámutattak a morphdb.hu határozószavaihoz rendelt kategorizációban, melyek kézzel történő javítása a modell alapján jelentős mértékben egyszerűsödött, hiszen a hasonló kategóriába tartozó szóalakok egymáshoz közel helyezkednek el a térben, így a csoportok egyszerre kategorizálhatók, illetve a hibásan kategorizált szavak kakukktojásként könnyen feltűnnek a hozzájuk hasonló viselkedést mutató, de másképp címkézett szavak között.

Érdekes megfigyelés a modellel kapcsolatban, hogy különböző távolságmetrikákkal számolva, illetve a korpusz különböző (pl. lemmatizált és elemzett vs. elemzetlen) változatain betanítva a modellt, láthatóan különböző szempontok dominálnak az osztályozásban. Ez egyben a Humor adatbázisban különböző egyébként nagyjából ekvivalens elemzésekkel kapcsolatos problémákra is rávilágított. Mikor a morphdb.hu határozószóként annotált lexémáit az (alapvetően a Humor elemző és a Szeged Korpusz felhasználásával) egyértelműsített és lemmatizált korpuszváltozatot használva klasztereztük, létrejött egy nagy méretű, grammatikai szempontból igen heterogén elemeket tartalmazó klaszter, amely – mint kiderült – olyan elemeket tartalmazott, amelyek a korpuszban főleg valamilyen mélyebb elemzéssel a szóalaktól különböző lemmára visszavezetve szerepeltek, de néhány elemzésük mégis atomi határozószói elemzést kapott (a tanítóanyag nem egységes annotációja miatt). Ezek a ritka elemzések azonban nem

[4] https://code.google.com/p/word2vec/

adtak megbízható disztribúciós modellt, így a furcsa klaszterben ezek a heterogén elemek jelentek meg, amelyekben leginkább az volt a közös, hogy az atomi határozószói elemzésük inkább ritka és zajszerű kivétel volt a másik részletesebb elemzésük mellett.

A disztribúciós elemzés hatásosan kimutatott egyéb annotációs anomáliákat is a korpuszban. Például jól elkülönülő klaszterekben gyűltek össze a Humor elemző által fel nem ismert, és így a PurePos tagger guessere által azonos módon hibásan elemzett és lemmatizált szavak (pl. azok, amelyekről mind tévesen vágta le az *-it* végződést, azt az *-i* képző tárgyragos alakjának vélve, vagy a téves anaforikus birtokos elemzés nyomán az *-é* levágásával, illetve a *-ba/be* végződés téves levágásával kapott hibás lemmák). A 3. táblázatban ilyen, illetve a korpuszban eleve hibásan írt vagy elválasztott és a tokenizáló által helyre nem állított szóalakokhoz tartozó klaszterek első néhány elemére láthatunk példát.

3. táblázat. Példák hibásan lemmatizált, tokenizált vagy hibásan írt kifejezésekre és a rendszer által azonosított hozzájuk hasonló szavak listájának első néhány eleme

pufidzsek	angolúl	kony	gyűrűj	sebti
rövidnac(43)	magyarúl(486)	nű(165)	királynőj(30)	Juteszem(33)
napszemcs(37)	németűl(132)	lyos(227)	Manass(104)	útköz(323)
szemcs(37)	francziául(25)	legha(17)	Hekat(91)	juteszem(94)
szmöty(45)	angolol(27)	komo(167)	oké-ok(412)	subscri(55)
zacs(170)	írül(95)	latos(283)	Lüzisztrat(21)	neszójjá(11)
suzuk(131)	mindenrõl(422)	legki(36)	juhéé(97)	kizom(25)
sap(374)	minderrõl(129)	csó(183)	jóskán(60)	akurvaélet(33)
törcs(11)	ilyenröl(58)	pontosab(59)	örüjj(35)	Egyfolyta(21)
kispolszk(41)	Amirõl(143)	nyolult(18)	Béb(774)	hébehó(25)
févör(8)	olyasmirôl(38)	kes(2114)	hoppár(189)	CsimbaWam(20)

A kötőszavak osztályozása a Humor adatbázisban és a morphdb.hu-ban részben különbözik. A Humor leírásban csak azok a szavak kaptak kötőszó címkét, amelyeknek a tagmondaton belüli disztribúciója egyértelműen kötőszószerű eloszlást mutat, és a pusztán a pragmatikai funkciójuk szempontjából tagmondatok közötti kötőelemként működő, de egyébként a tagmondatbeli elhelyezkedésük szempontjából határozószószerű viselkedést mutató szavak következetesen határozószóként vannak osztályozva. Ezzel szemben a morphdb.hu-ban ezek egy része is kötőszó címkével szerepel. Az egységesített címkézési rendszerben ezeket a határozószók egy alosztályaként címkézzük. Ugyancsak nem kötőszóként szerepelnek az egyébként a kötőszók eloszlásának megfelelő viselkedést muta-

tó vonatkozó névmások, hanem ezeket a főkategóriájukat megtartva vonatkozó névmásként alkategorizáltuk. A morphdb.hu nyelvtan áttekintése után a morphdb.hu lexikonban szereplő elemek jegyeit is automatikusan át tudjuk vinni a Humor lexikonba. Ugyanakkor kézi ellenőrzést is igényel a folyamat, mert a megadott jegyek sok esetben tévesek, vagy hiányoznak. Például sok helyesírási anomáliát tartalmazó szótő nincs ilyenként megjelölve.

4.3. A keretrendszer

A létrejött forrásadatbázis nyelvész szakértő számára olvasható, értelmezhető és plain text editorral szerkeszthető. Olyan kiegészítő alkalmazás fejlesztésére is sor kerül, amely a tőtár bővítését és megváltoztatását számítógépes nyelvészeti szakértelem nélkül is lehetővé teszi.

A második réteget képező forrásadatbázis-konverter a forrásadatbázis tőtárából *lexc* [1] lexikont állít elő, melyet a HFST keretrendszerben [4] implementált harmadik réteg használ fel.

A létrejövő keretrendszer lehetővé teszi a forrásadatbázisban specifikált információ alapján testre szabható domén- és regiszterspecifikus elemzők előállítását. Kimeneti kódkészletként bármely, magyar nyelvre elterjedt kódrendszer (KR, Humor, MSD) választható, illetve a már meglévő annotált korpuszokkal való kompatibilitás is biztosított. A lemmatizálás során a tő részét képező toldalékolás, illetve az elemzés során figyelembe vett morfofonológiai jellemzők konfigurálhatók.

5. Konklúzió

A cikkben egy olyan készülő nyílt forráskódú magyar morfológiai elemző létrehozására irányuló fejlesztést mutattunk be, amely több korábban készült magyar számítógépes morfológiai leírást harmonizálva és egyesítve, illetve azt további lexikai tudással kiegészítve terveink szerint minden korábbinál teljesebb és pontosabb eszköz lesz. Folyamatban van a rendszer alapjául szolgáló Humor és a morphdb.hu magyar morfológiai leírások harmonizációja és egyesítése. A lexikon ellenőrzéséhez és további lexikai jegyek félautomatikus felvételéhez folytonos vektortéralapú disztribúciós modelleket és automatikus hierarchikus klaszterezőalgoritmust használunk, amelyek igen hatékony eszköznek bizonyultak az elvégzendő lexikográfiai munka támogatásához.

Hivatkozások

1. Beesley, K., Karttunen, L.: Finite State Morphology. No. 1 in CSLI studies in computational linguistics: Center for the Study of Language and Information, CSLI Publications (2003)
2. Huldén, M.: Fast approximate string matching with finite automata. Procesamiento del Lenguaje Natural 43, 57–64 (2009)
3. Laczkó, K., Mártonfi, A.: Helyesírás. A magyar nyelv kézikönyvtára, Osiris (2004)

4. Lindén, K., Silfverberg, M., Pirinen, T.A.: Hfst tools for morphology - an efficient open-source package for construction of morphological analyzers. In: Mahlow, C., Piotrowski, M. (eds.) SFCM. Communications in Computer and Information Science, vol. 41, pp. 28–47. Springer (2009)
5. Novák, A.: Milyen a jó Humor? In: I. Magyar Számítógépes Nyelvészeti Konferencia. pp. 138–144. SZTE, Szeged (2003)
6. Novák, A.: A new form of humor – mapping constraint-based computational morphologies to a finite-state representation. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1068–1073. European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014), aCL Anthology Identifier: L14-1207
7. Prószéky, G., Kis, B.: A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 261–268. ACL '99, Association for Computational Linguistics, Stroudsburg, PA, USA (1999)
8. Siklósi, B., Novák, A.: Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra. In: Tanács, A., Varga, V., Vincze, V. (eds.) XII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 3–14. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2016)
9. Trón, V., Halácsy, P., Rebrus, P., Rung, A., Vajda, P., Simon, E.: E.: Morphdb.hu: Hungarian lexical database and morphological grammar. In: Proceedings of LREC 2006. pp. 1670–1673 (2006)